Scraping meaning from technical information:
Semantic mining of chemical information from non-chemistry disciplines

Aaron Pital
Stockton Group, Georgia Institute of Technology, School of Chemistry and Biochemistry

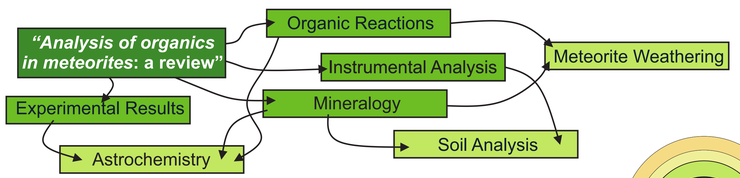
Georgia Tech

Information is being created and shared faster than humans can read and assimilate it

We're forced to use strategies to cut down the volume of information into manageable chunks somehow. There are a few strategies: dimensionality reduction[1], network analysis[2-5], neural-network derived keyword handling [6], etc. A major issue of scaling still remains. It is intractable because most systems of information are complex at every scale, and attempting to look at the bigger picture obscures the fine structure where innovation is made, and zooming in leads to navel gazing. Many search engines (civilian and academic) get around this to some extent by amplifying what the community finds important through measuring clicks, citations, or mentions. This can lead to echo chambers, however, and in science can exacerbate inequalities

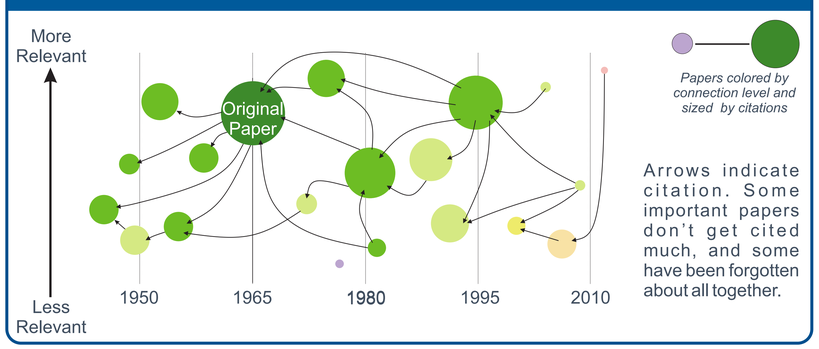
Your search returns 1,460,000 results in 0.84 sec. Are the top results important, or just popular?

A colleague mentions an unfamiliar concept and you want to learn the basics. How many papers and books do you need to read to get the required background knowledge and basics concepts?



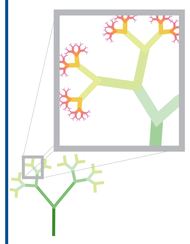
Every new concept in the sciences is built on the shoulders of prior work. A cursory understanding of new work often requires some understanding of the background as well. Knowing what is important and what can be skipped is difficult without a formal background in that subject.

Your boss assigns you to learn about a famous reaction and find out how to implement it in your lab and what other groups used it for in the last 50 years. How do you find relevant work?

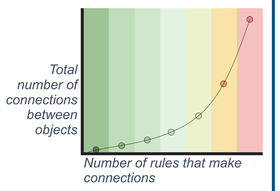


How to make sense of large sets of information?

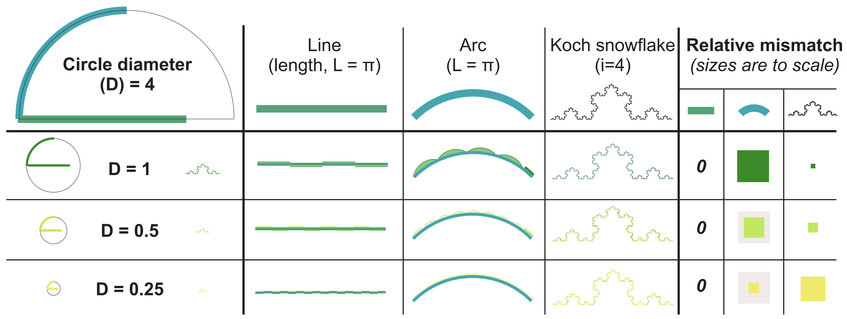
Knowledge systems branch at every scale; they're fractal



Every new keyword match, citation link, or shared authorship adds more documents to the set of all documents considered no matter what kind of analysis is being done. Measuring connectedness in these types of systems is difficult because fractals don't behave like normal systems when you try to measure them. **Finding connections between technical documents gets harder the more documents that are considered.**



Measurements in fractal systems blow up quickly and make smaller features hard to see



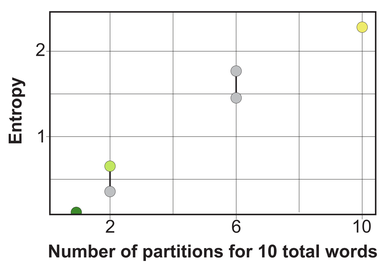
Using entropies (in the information theory sense) is a solution

10 words, 10 partitions
"Every, amino, acid, can, be, found, in, meteoritic, samples, today "

10 words, 2 partitions
"Amino, Acids, Amino, Acids, Amino, Acids, Amino, Acids, Amino, Acids"

10 words, 1 partition
"Meteorite, Meteorite, Meteorite, Meteorite, Meteorite, Meteorite, Meteorite, Meteorite, Meteorite, Meteorite"

Entropy (S) = -Σ *(word_probability)* **ln(***word_probability***)**



How are entropies used on real data?

Set of papers, documents, etc.
Important Technical Paper 1

Make a dicionary from documents (preferably with software)
Read each word
Check to see if it's new
Add new words to dictionary

Process documents w/ dictionary (preferably with software)
Count each word's occurrence
Calculate entropy as above
Assign weights between papers

In this work:

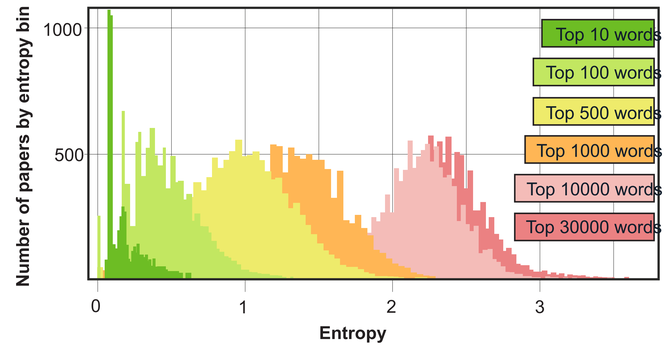
Set of 11,500 papers on meteorites, organic content, and papers linked to famous papers by citations or authors

Processed in Python 3.6
NLTK, Gensim, Bokeh

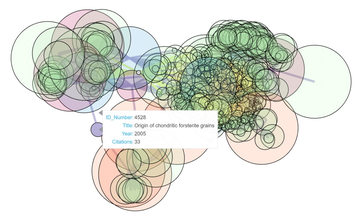
Results

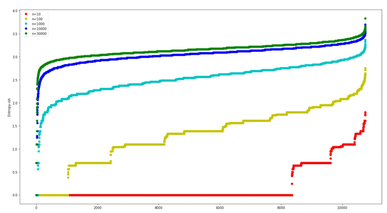
Information entropies behave like physical entropies

Entropies for each of the 11,500 documents was calculated using dictionaries of various lengths. The more words considered, the larger the range of possible entropies with most falling a normal distribution. This is exactly the statistical behavior of ideal gasses if instead of the 'ruler' of dictionary length one uses the temperature of the gas at equilibrium.



Because the response is statistical, deviations from the expected statistics highlight that something interesting is happening. By comparing not the values for each paper themselves but how the change in each paper's entropy value as different 'ruler' (dictionary) lengths are used, **we can tell which papers are most connected to which other papers' word use, how strong that connection is, and which particular words are responsible.** This analysis is not a model and can be saved and used again. **When new documents are considered, the old analyses don't have to be redone, and old results can directly be used in new computations.**

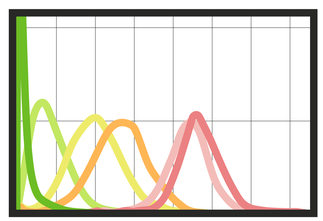




Future Work

More papers will be added to the analysis set. This will allow for more nuanced connection data between technical documents. Reference analyses on textbooks and reviews are being done now.

Authorship pages are being scraped and analyzed for human connections to published works. One goal of these analyses is to highlight collaborations between scientists, industry partners, and government that are mutually beneficial and revolve around the same technical and policy areas but which are difficult to identify because of siloed interests and non-overlapping professional circles.



References

[1] Astrobio. (2015), 15(7), 587-600.
[2] Amer. Miner. (2008), 93, 1693-1720.
[3] Astrobio. (2016),16(2), 181-97.
[4] J.Mol.Evol. (2016), 83(1-2), 1-11.

Acknowledgements

I'd like to thank the Stockton Lab for their continued support in all ways big and small, and Amanda Stockton for her outstanding mentorship.